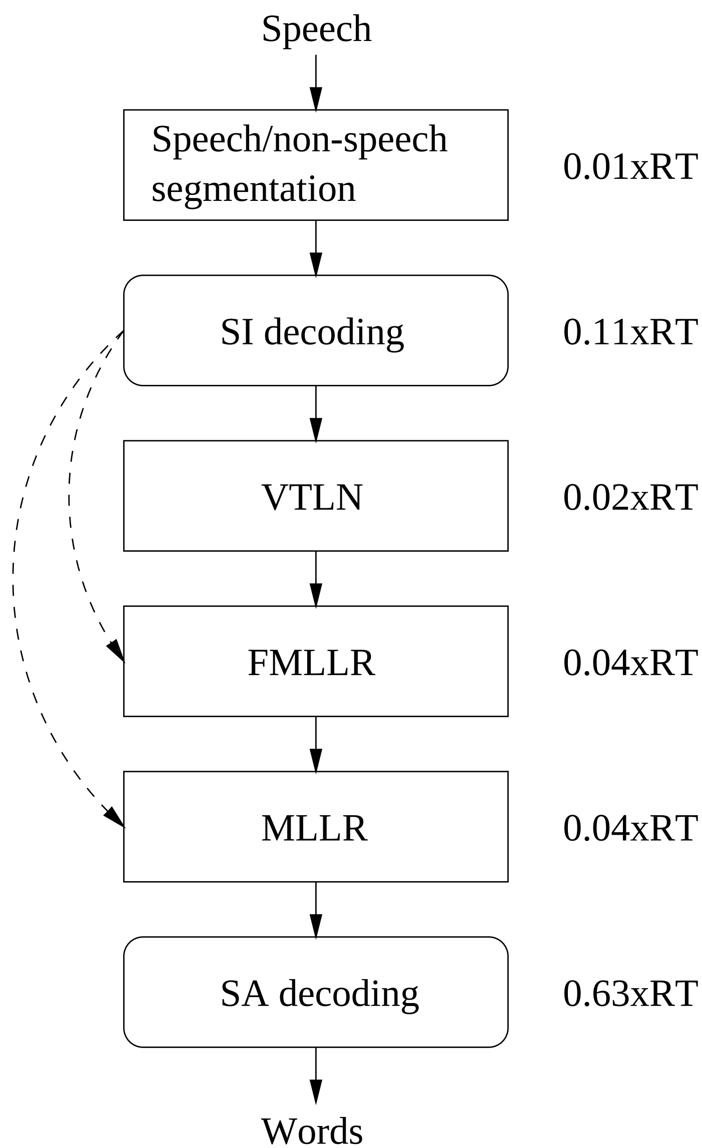
The IBM 2003 1xRT speech-to-text system

George Saon, Geoffrey Zweig, Brian Kingsbury and Lidia Mangu

Outline

- System diagram
- Speech/non-speech segmentation
- Front-end processing
- Acoustic models
- Speaker compensation
- Static graph decoding
- Conclusion

System diagram



Speech/non-speech segmentation

Viterbi decoding for a two word vocabulary

- Each type of segment is modeled by a 5-state HMM
- Segment insertion penalty controls number and duration of segments
- Speech (resp. non-speech) GMMs tied across all states in a model
- GMMs obtained by bottom-up clustering of the SI Gaussians (123 for speech and 12 for silence)
- Hypothesized speech segments extended by additional 30 frames
- Overlapping segments are merged together

Segmentation performance

- Number of segments:

Reference	AT&T*	IBM
9050	9012	6661

- Word error rate:

	Reference	AT&T	IBM
SI decoding	49.3%	49.5%	49.5%
SA decoding	28.7%	29.2%	29.0%

- Speed: 0.008xRT (3m3s)

*Courtesy of Andrej Ljolje

Front-end processing

Two types of features:

- 24-dim MFCCs for segmentation and speaker independent decoding
- 13-dim VTL-warped PLP cepstra for speaker adapted decoding

Common characteristics:

- 25ms Hamming window, 10ms shift
- Spectral flooring by adding 1 bit prior to the Mel binning
- Periodogram averaging (Welch smoothing)
- Every nine consecutive frames are concatenated and projected down to 60-dim through LDA+MLLT

Acoustic models

- Phonetic questions within an 11-phone window with left cross-word acoustic context only
- Leaves of the decision tree are modeled by at most 128 diagonal covariance Gaussians
- Number of Gaussians determined using BIC

Number of	SI	SA
leaves	4.0K	4.6K
Gaussians	168K	158K

- SAT models trained through implicit-lattice MMIE [IBM RT'02]
- Training data: 247 hours of Switchboard, 18 hours of Callhome and 18 hours of Switchboard cellular

Speaker compensation

1. Alignment-based VTLN:
   - 21 warp scales allowing for a $\pm 20\%$ stretching of the frequency axis
   - Selectively score vowels
   - Jacobian compensation
   - Uses at most 60 seconds of test data per speaker

2. Alignment-based FMLLR (1 transform):
   - Maps the VTL-warped test data to a canonical SAT feature space
   - Statistics accumulated in single precision

3. Alignment-based MLLR (1 transform):
   - Statistics accumulated in single precision (necessary to scale means by standard deviations to avoid overflow)

All compensation steps use the Intel MKL library extensively

Speaker compensation performance

- Runtimes and RTFs:

	Runtime	RTF
VTLN	5m10	0.013xRT
FMLLR	13m59	0.038xRT
MLLR	17m39	0.048xRT

- Word error rates:

	RT'02	RT'03
SI	50.3%	49.5%
VTLN	34.1%	32.9%
FMLLR	30.6%	29.7%
MLLR	30.1%	29.0%

- Effect of improved SI:

	1xRT	2xRT
SI	49.5%	37.1%
MLLR	29.0%	28.7%

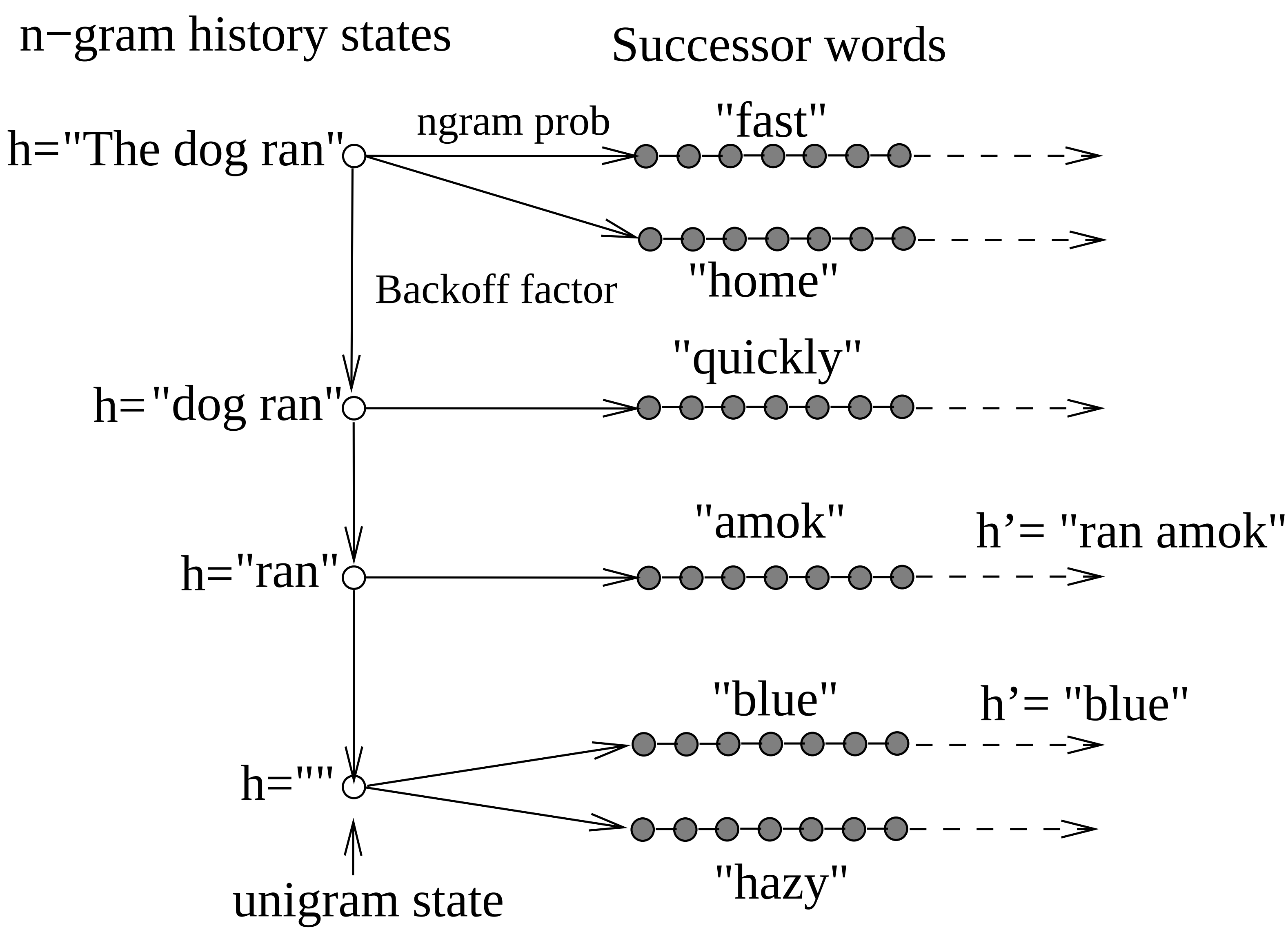
Static graph decoding

- SI and SA decodings operate on static FSM graphs
- Backoff LM expansion at the HMM level (2-gram for SI, 4-gram for SA)
- Arc minimization for cross-word context [Zweig, Yvon & Saon'02]
- State determinization and minimization [Mohri, Perreira & Riley'00]

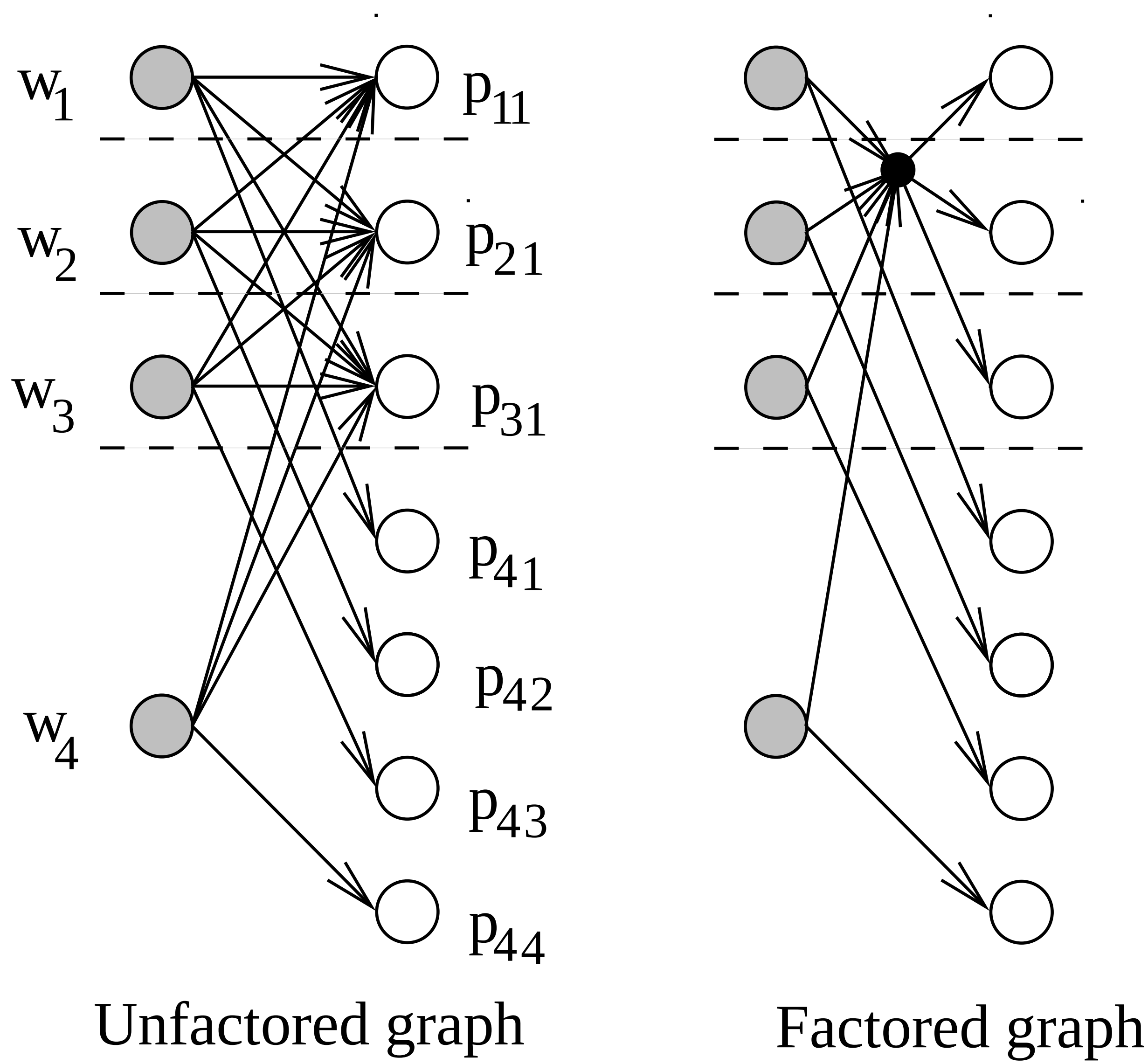
Number of	SI	SA
ngrams	0.2M	3.3M
states	0.6M	9.6M
arcs	1.7M	23.9M

- LM training data: 3M words Switchboard, 59M words web scripts UW, 3M words BN, 7M words English Gigaword and 1M words from BBN

LM expansion



Cross-word arc minimization

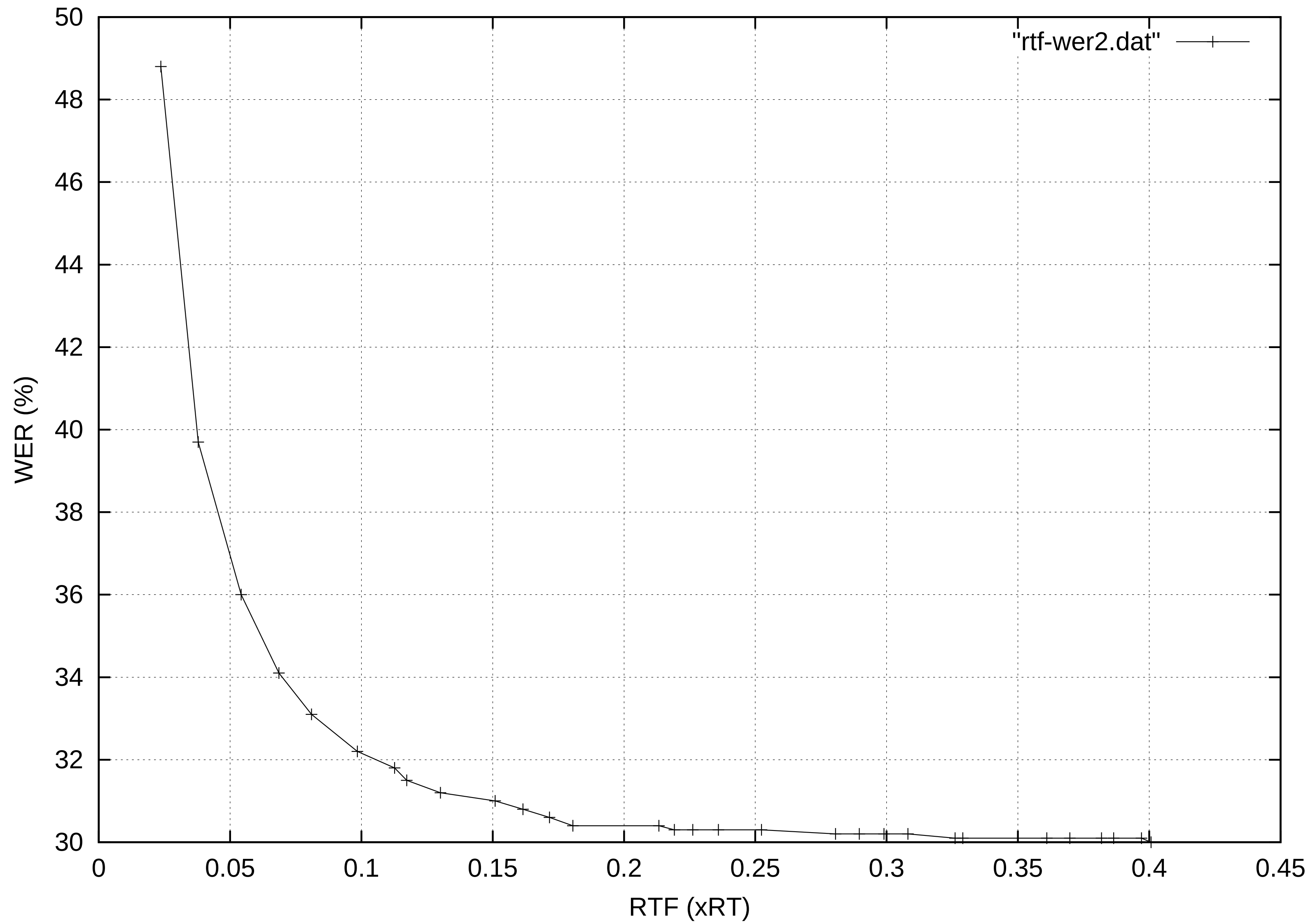


Decoder characteristics

- State (rank) pruning as opposed to beam pruning (500 active states/frame for SI, 3500 active states/frame for SA)
- Hierarchical Gaussian evaluation decoupled from search
  - Components clustered to 2048 Gaussians. For each frame, evaluate only the Gaussians which map to the top N clusters. (N=20 for SI and N=110 for SA)
  - Streaming SIMD extension 2 instructions of the Pentium 4 processor
  - Gaussians sorted by top-level cluster
- Handling of layers of null states (observations emitted on states not arcs)
- Search errors due to pruning:

SI	49.5%/0.107xRT	40.4%/1.2xRT
SA	29.0%/0.628xRT	28.4%/1.1xRT

RT'02 WER-RTF performance (search only)



Conclusion

- Two-pass decoding strategy with 3 adaptation passes inbetween
- Static graph decoding is the only way for an accurate 1xRT system
- Single transform adaptation is limited
- No consensus, no rover
- 0.2xRT loss due to I/O